Evaluating x-ray microanalysis phase maps using principal component analysis.

Ben Buse and Stuart Kearns

Abstract

Automated phase maps are an important tool for characterising samples but data quality must be evaluated. Common options include overlay phases on BSE images and phase composition averages and standard deviations. Both these methods have major limitations. We propose two methods of evaluation involving principal component analysis. First, a RGB composite image of the first three principal components, which comprise the majority of chemical variation, and provides a good reference against which phase maps can be compared. Advantages over a BSE image include discriminating between similar mean atomic number phases and sensitivity across the entire range of mean atomic numbers present in a sample. Second, principal component maps for identified phases, to examine for chemical variation within phases. This ensures the identification of unclassified phases and provides the analyst with information regarding the chemical heterogeneity of phases (e.g. chemical zoning within a mineral, or mineral chemistry changing across an alteration zone). Spatial information permits a good understanding of heterogeneity within a phase and allows analytical artefacts to be easily identified. These methods of evaluation were tested on a complex geological sample. K-means clustering and K-nearest neighbour algorithms were used for phase classification, with the evaluation methods demonstrating their limitations.

Key words

Electron probe microanalysis, scanning electron microscopy, phase mapping, k-means clustering, wavelength dispersive spectroscopy (WDS), energy dispersive spectroscopy (EDS), elemental maps, clustering

Introduction

Automated phase mapping is a widely available tool within software suites for energy dispersive spectrometers (EDS) on scanning electron microscopes (SEM) and is now available for electron probe microanalysis (EPMA) using wavelength dispersive spectrometers (WDS). Phase mapping uses multi-dimensional data (spatially defined element intensities or concentrations) and identifies chemically distinct phases to give their spatial distribution. The automated algorithms included in the instrument software make this process straightforward for the operator (e.g. a form of principal component analysis using rotation (Kotula et. al 2003), clustering algorithm in Oxford Instruments AutoPhaseMap software (Statham et al. 2013), K-means clustering in Probe for Epma software (www.probesoftware.com) and hierarchical cluster analysis in JEOL EPMA software (Mori et al. 2017)).

It is difficult to assess the quality of a phase classification, Munch et al. (2015) demonstrate some of the limitations of phase algorithms and suggest the need for expert review. Common options to evaluate phase classifications include overlay on BSE images and phase composition averages and standard deviations. Liebske (2015) provides an improvement in the open source package iSpectra which allows overlays on principal component maps or RGB elemental maps. Phase composition averages are often difficult to interpret being affected by convoluted pixels at grain boundaries, where the limits of analytical resolution results in measured intensities consisting of a convolution of adjacent phases. van Hoek et al. (2011) and Liebske (2015) showed how these ‘bad’ pixels can be eroded to give phase composition averages reflecting true compositions but this processing is not

available in most phase mapping packages. Algorithms can be independently verified for reference samples against methods such as manual thresholding (Maloy & Treiman 2007) or EBSD phase maps (Statham et al. 2013), but this does not ensure the algorithm works correctly for all samples and operating conditions (Munch et al. 2015)

In this study k-means clustering and k-nearest neighbour (KNN) algorithms are used to demonstrate some of the problems and show how, irrespective of the phase mapping algorithm, principal component analysis (PCA) can be used to assess the quality of phase classification. PCA assigns new dimensions which capture the variability of the dataset; the first dimension corresponds to maximum variance; each subsequent dimension is orthogonal to the previous and captures the maximum remaining variance (Tan et al. 2006). The merit of this technique is it provides an unbiased method of reducing multiple dimensional systems (e.g. 10 chemical elements) to a small number of dimensions which can be visualised graphically. PCA and various refinements are commonly used in the generation of phase maps (e.g. Kotula et al. 2003, Parish 2011), here we demonstrate their strength in evaluation of phase maps.

This study uses quantitative maps for the phase classification. Quantitative maps provide significant advantages over raw count maps, allowing the interrogation of phase data and importantly average phase compositions to be extracted.

Methods and materials

A complex sample was selected with multiple minerals of varying abundance and finely intergrown minerals at or below the limits of analytical resolution (controlled by accelerating voltage and pixel/step size) (see Figure 1c). The sample is a metamorphosed basalt xenolith within a kimberlite (for details see Buse et al. 2010). The area mapped extends from the kimberlite into the basalt xenolith (Figure 1a) with alteration of the basalt most intense adjacent to the kimberlite. Quantitative element maps were collected using 5 WDS on a JEOL 8530F EPMA at the University of Bristol. Elements collected were Si, Na, Ca, Fe and Ti in the first pass and Mg, Al, K, Mn in the second pass. The operating conditions were 20 kV accelerating voltage, 40 nA beam current, 10 millisecond dwell time and a 5 µm step size. The quantitative element maps are combined into a single array (X coordinate, Y coordinate, Si, Na, Ca, Fe, Ti, Mg, Al, K, Mn) for processing.

The phase maps were generated and evaluated using R (General Public License software for statistical computing), which includes k-means clustering, k-nearest neighbour and PCA packages.

K-means clustering requires the initial cluster centres to be specified or determined randomly for a given number of clusters. Data is classified through a series of iterative loops in which all the points (pixels) are assigned to the nearest cluster centre (centroid) and the centroid position is updated. Here K-means clustering was run in three variants: (1) using randomly assigned initial cluster centres for 15 clusters; (2) using specified cluster centres for discrete phases identified from the Red-Green-Blue (RGB) composite image of principal components. 9 discrete phases were identified (Figure 1b and Table 1). For each discrete phase a single area composition was extracted from the element maps; (3) using maximum element intensities were used as the initial cluster centres.

K-nearest neighbour requires a reference dataset against which the pixel compositions are checked. The reference dataset consisted of the compositions of the 9 discrete phases (Figure 1b and Table 1) selected for K-means specified cluster centres. A normal distribution using the measured standard deviation was applied to each composition to present a range of compositions for each phase. For each pixel the 10 closest reference values in chemical space were examined, a pixel was assigned to a phase if at least 7 of the reference values belonged to the same phase, otherwise it was rejected.

For PCA, the dataset was centred in principle component space so that the mean of each principle component is 0 rather than the mean of the compositional data. This ensures that the first principle component is not dominated by the position of the dataset with respect to the origin (Jolliffe 2002). The dataset was not scaled (a covariance matrix was used); scaling (a correlation matrix) gives equal weight to the variance of each component (here chemical element) and is important where components of different units (e.g. length, weight etc.) are being analysed (Jolliffe 2002). It is not desirable here, because the units of each of component (wt.%) are the same. By not scaling, the variance is dominated by major element variance rather than giving trace elements equal weight. This is desirable because the phase separation is based on major constituents and noise dominates the trace element signals (van den Berg et al. 2006). PCA was conducted on the entire dataset of multiple element intensities for the whole map to produce RGB composite images of the first three principal components. PCA was also conducted on subsets of the data, consisting of the element intensity pixels for a single phase to produce phase-PC (principal component) maps and scatter graphs.

Results

Red-Green-Blue Principal Component (RGB-PC) images

Figure 2 compares several phase mapping algorithms to a backscattered electron (BSE) image and an RGB composite image consisting of the first three principal components (PC), derived from principal component analysis. The RGB-PC image provides a good tool to assess phase maps; similar to the use of BSE images in some phase analysis software (e.g. Thermo Scientific NSS permits overlays of phases onto a BSE image). Figure 2b shows that the RGB-PC image provides a greater phase separation than the false-colour BSE image (Figure 2e). The BSE image has difficultly both separating phases with similar mean atomic number (e.g. on Fig 2e colours of bultfonteinite [yellow-green], clinopyroxence [green-blue], and serpentine [blue] overlap; see Table 1 for mineral compositions) and covering the range of mean atomic numbers present in the image. The high atomic number phases ilmenite, perovskite and barite cannot be separated (red on Fig 1e) with the detector brightness and contrast set for sensitivity at the low mean atomic numbers. BSE images are more sensitive to topography than most x-ray intensities. The RGB-PC image is derived from the same dataset (x-ray intensities) as the phase maps. BSE images can still make a contribution in checking phase maps; dependant on mean atomic number they can identify variations not measured by x-ray intensities (e.g. H2O in normalised EDS data; Munch et al. 2015) – in the case of figure 2 only the BSE image differentiates between barite and holes – with sulphur and barium not measured. Whilst the phase mapping system, using only WDS data for the measured elements, cannot be expected to differentiate between the two, a comparison with the BSE image alerts the analyst. The RGB-PC image extends the evaluation tools suggested by Liebske (2015) and accounts for most of the chemical variation within the sample.

The RGB-PC image provides a reference for a visual assessment of the phase maps. The number of phases and textural features (e.g. shape of grains) can be checked. In the examples given, K-means with 15 random clusters (Fig. 2a) subdivided hydrogarnet and serpentine into numerous phases (on Fig. 2a hydrogarnet is orange, dark-red and black, and serpentine is blue, pink and violet; see also Table 2). The other phase maps (Fig. 2c,d & f) provide a close match to the RGB-PC image. In addition K-means with 15 random clusters (Fig. 2a) identifies ilmenite and perovskite as a single “oxide” phase, which from RGB-PC image can be seen to consist of chemically distinct phases. The other phase maps correctly separate this “oxide” phase into ilmenite and perovskite. This distinction is critical for interpreting the sample, with ilmenite absence from the margin of the basalt xenolith as a result of alteration penetrating into the basalt from the kimberlite (Buse et al. 2010).

Serpentine and bultfonteinite form fine intergrowths below analytical resolution; on the RGB-PC image (Fig. 2b & h) this intergrowth form a distinct phase (dark purple distinct from the bright pink of serpentine). This phase is well characterised using the KNN algorithm, which on Fig 2i in comparison with the RGB-PC image can be seen to faithfully reproduce the serpentine, serpentine-bultfonteinite intergrowth, and hydrogarnet phases. Again K-means with 15 random clusters can be seen to split the phases into many subdivisions (serpentine-bultfonteinite intergrowth is split into dark brown and grey phases). The K-means, using specified clusters or maximum intensity, struggles in the classification of serpentine and serpentine-bultfonteinite intergrowth. Serpentine and serpentine-bultfonteinite intergrowth are under-represented, whilst a mixed serpentine phase (pink on Fig 2l & j) and bultfonteinite are over represented (see black arrows on Fig2l in comparison with Fig2i and Fig 2h).

The main distinction between KNN and K-means using specified clusters or maximum intensity, is that in the latter the cluster centre can shift during the iterative process. The result of this is shown in comparison with Table 1 where phase 2 has shifted and now represents an additional mixed serpentine phase, which diminishes the serpentine-bultfonteinite intergrowth phase (pink phase; Fig2l & j) and misclassifies convoluted boundary pixels. The latter is seen in Figure 3c where the black arrow identifies pink boundary pixels, a convolution of serpentine and hydrogarnet, not visible on either the RGB-PC image (Fig 3b) nor the KNN phase classification (Fig 3a). Another example of iterative shifting of the cluster centre is phase 9 in Table 1 (lilac phase on phase maps) which has shifted so that the phase includes both the initial apatite (see Fig 3c), mixed phases (see Fig 3c & 2l circles) and pits and barite (Fig 2 d & f).

Phase analysis software commonly report phase composition averages and standard deviations. Table 2 gives the values for K-means with 15 random clusters. The values are difficult to interpret as averages may differ significantly from the true composition. Table 3 gives the composition of clinopyroxene extracted from several pixels within a single clinopyroxene crystal, here the composition closely matches stoichiometry. Poor phase composition averages are often the product of analytical resolution (see van Hoek et al. 2011, Liebske 2015) as pixels at the margins of grains can have convoluted x-ray intensities of multiple phases. The phase may also include bad pixels where topography gives poor results again skewing the phase average. Table 2 also shows the problems of identifying mixed phases resulting from small grains dominated by convoluted pixels of boundaries or finely intergrown phases. To correctly identify phases, comparison with a RGB-PC image can be of considerable help.

Phase Principal Component (phase-PC) maps

Phase-PC maps provide a useful tool to assess the homogeneity of each phase and determine whether it includes multiple phases which the algorithm has failed to discriminate. Figure 4a shows a phase-PC map of the “oxide” phase generated from the k-means algorithm using 15 random clusters. The phase-PC map clearly distinguishes ilmenite (orange) from perovskite (purple). The spatial separation and the magnitude of variance provides strong evidence for two distinct phases.

Phase-PC maps for perovskite and ilmenite, which the KNN algorithm correctly identifies as two separate phases, shows the perovskite to be relatively homogenous whilst ilmenite contains two spatially and chemically distinct phases. Both the perovskite and ilmenite phases include some chemical variation from convolution at the margins of grains. Using PC-1 ilmenite can be subdivided into two compositions (Table 4). The small purple grains in the kimberlite (identified on Fig. 4c) represent Fe-Ti-Mg spinel are distinct from the ilmenite within the basalt xenolith.

The PC-1 Phase-PC map shows this distinction less clearly for ilmenite identified using k-means with specified clusters. The ilmenite data contains more scatter than observed for the KNN ilmenite phase. This is consistent with k-means not rejecting any 'bad' pixels, unlike KNN. For the low abundance phase this scatter has significant influence and results in the rotation of the principal components (see Figures 4g and 4h). In this case the PC-2 Phase-PC map most clearly distinguishes ilmenite from Fe-Ti-Mg Spinel, whereas PC-1 shows variations within grains suggestive of distinguishing convoluted pixels.

Figure 5 shows Phase-PC maps for clinopyroxene, serpentine and bultfonteinite, all of which are relatively homogenous. A comparison with the PC scatter graphs illustrates the benefit of spatial information. Both clinopyroxene and serpentine show small PC variations. On the maps it is evident that for clinopyroxene this is uniformly distributed whereas for serpentine it varies between the kimberlite and the xenolith. Variation within the clinopyroxene probably relates to pixel convolution although could relate to chemical zonation within the clinopyroxene. Variations within the serpentine suggest the chemistry of the serpentine differs spatially. The average compositions (Table 5) are inaccurate and difficult to interpret possibly due to convoluted pixels as discussed above. Beam damage may also add to the reduced data quality. However, variations in Al, Si, Mg and Fe, with Al enriched in the kimberlite are clearly apparent. The presence of spatial variations provides important information about the sample, which should prompt further detailed investigations to understand the cause. Element maps extracted for the serpentine phase (Figure 6) confirm the variations suggested by the average compositions (Table 5) with Al substituting for Si and Mg for Fe.

Figures 5 c-d compare bultfonteinite from K-means using 15 clusters and from K-means specified clusters. The difference can be explained as K-means specified clusters has fewer clusters resulting in the bultfonteinite phase being less tightly constrained and containing marginal data. This incorporation of marginal data is shown in Fig. 5d where the centre of the grains (purple) corresponds closely to Fig 5c, whereas the rest of the data (orange) consists of increasingly mixed compositions excluded from the more tightly constrained cluster of K-means using 15 clusters.

Discussion

Phase classification is complex and subject to the limitations of the algorithm used. PCA provides a method of evaluating the quality of a given phase classification method. PCA is used in reference to phase maps and element maps: checking that the phase maps represent the variation identified in the RGB-PC image and checking for any variation within an individual phase. In the latter case, variation within a phase is explained by the elemental data extracted for the discrete variations in PC identified (e.g. Table 4 where extracted compositions allowed Fe-Ti-Mg spinel to be identified within the ilmenite phase). This use of PCA in reference to phase maps and element maps avoids the difficulties associated with interpreting principal components from their component weights (see Kotula et al. 2003). PCA requires an orthogonal arrangement of components which may not correspond to data variation (Kotula et al. 2003) as shown in the phase-PC maps and scatter graphs in Figure 4 d, e and h. This problem is mitigated in the case of RGB-PC images for it is a composite of 3 principal components. In some cases the orthogonal requirement can obscure variations in phase-PC maps suggesting in these cases phase-PC maps for each principal component are required, or possibly scatter graphs or RGB-PC images for the phase. Improvements might be possible through rotating PC or by removing the orthogonal constraint. Regardless the phase-PC maps show the value of this or similar techniques in identifying variations in multi-dimensional space within individual phases and displaying their spatial component.

The data presented shows how PCA can be used to identify incorrect phase classifications; here exposing the limitations of K-means clustering using random clusters. K-means works best for phases of similar abundance, which form spherical clusters in chemical space and where the initial allocated centres reflect phase distribution (Tan et al. 2006). Both the RGB-PC image and the phase PC maps identify the "oxide" phase and the phase PC maps show the "oxide" phase to actually consist of perovskite and ilmenite. Due to their low abundance, these phases are not distinguished using randomly allocated cluster centres, which only subdivides more abundant phases (Fig. 7, see also Munch et al 2015).

Specifying initial cluster centers, either by identifying phases beforehand (K-means specified clusters, Fig 2d, see also Munch et al. 2015) or by using maximum element intensities (K-means max element, Fig 2f), to a large extent overcomes these limitations by ensuring the initial cluster centers represent phase distribution. The use of maximum element intensities does not require prior knowledge of phases but requires the number of phases to match the number of elements and for phases to be discriminated to a large extent by a particular element. A variant on this is using the KNN algorithm which does not iteratively shift cluster centres and allows pixels to be rejected (not classified). KNN gives consistent results for spatially distinct regions of the same rock sample; the absence of iteration means it is largely unaffected by the absence of a phase within an individual map. The danger with these methods is, in the case of specifying phases, not all the phases present in the sample may have been identified, and in the case of maximum element intensities, there may be more phases than elements. In these cases, the phase PC maps work well at identifying aggregate phases in which discrete phases have been classified together – as shown in the case of the "oxide" phase, and also the Fe-Ti-Mg spinel phase which was identified subsequent to phase analysis.

An alternative approach to overcoming the tendency for k-means to subdivide high abundance phases before distinguishing low abundance phases, is a set of criteria which recombine phases if certain thresholds are exceeded (Munch et al. 2015, Statham et al. 2013). With this approach, similar to using maximum element intensity, prior knowledge of phases is not required. However, it is still important to evaluate the output classification (Munch et al. 2015).

Convoluted pixels cause many problems in phase analysis and algorithms must ensure they are not assigned to distinct "boundary" phases. To identify true phase compositions these pixels should be rejected (van Hoek (2011), Liebske 2015) but for phase abundance, spatial distribution and textural shape they must be considered (Liebske 2015). For the KNN algorithm it is important that the phases within the reference dataset correspond approximately to the phases present in the sample. If the number of phases in the reference dataset greatly exceeds that in the sample, there is a high probability that the convoluted boundary phase pixels will have a composition similar to a phase in the reference dataset and be misidentified. When evaluating phases it is important to be able to distinguish between variance due to convoluted pixels and actual variation due to chemical variation within a phase or the presence of multiple phases. The example of Phase-PC maps for serpentine and bultfonteinite (Figure 5 b-c) show the importance of spatial information in making this assessment.

Principal component maps demonstrate how the generation of a phase map need not be the end of the process. New phases may be identified allowing the initial phase map to be revised. A phase could be subdivided based on principal components and its chemistry extracted or phase map algorithms could be rerun with an additional specified cluster. Where Phase-PC maps suggest variations within phases, e.g. as shown in the compositional difference between serpentine in the kimberlite and the basalt xenolith, the user can further investigate thus improving the sample characterisation.

Conclusions

In agreement with other work (e.g. Munch et al. 2015) the data presented illustrates the need for phase maps to be subjected to critical analysis, exposing any limitations of the algorithm or operating conditions resulting in incorrect classification. The performance of phase algorithms will vary depending on the sample (Munch et al. 2015) and the input parameters (the number of phases for K-means using random clusters; the phases specified for KNN and K-means using specified clusters), making it important to check the data has been correctly classified. Principal component maps provide an easy solution to evaluate phase classification. RGB PC images provide a good visual reference for checking phase maps, more clearly discriminating between phases than BSE images. Phase PC maps provide a good method of assessing variation within phases and identifying unclassified phases with the spatial information important for discriminating real chemical variance from convoluted pixels. The role of an operator in checking phase maps introduces subjectivity but the provision of spatial information allows the operator to make high-quality decisions as to the nature of variance, resulting in robust sample characterisation. This process of evaluation of phase maps allows further refinement and can provide additional information about a sample prompting further investigation.

KNN is potentially a very useful method of phase classification for geological samples, where the analysist is familiar with the possible phases within the rock sample. It produces consistent results similar to manual thresholding (e.g. Muir et al. 2012). K-means with specified clusters produces similar results but is more affected by the absence of a particular phase, when shifting from area to area within or between rock samples.

Acknowledgements

The authors would like to thank the reviewers for their insightful comments and revisions which improved the manuscript.

References

Buse, B., Schumacher, J.C., Sparks, R.S.J. & Field, M. (2010). Growth of bultfonteinite and hydrogarnet in metasomatized basalt xenoliths in the B/K9 kimberlite, Damtshaa, Botswana: insights into hydrothermal metamorphism in kimberlite pipes. *Contrib Mineral Petrol* **160**, 533-550

Jolliffe, I.T. (2002). Principal Component Analysis Second Edition. New York, USA: Springer-Verlag

Kotula, P.G., Keenan, M.R. & Michael, J.R. (2003). Automated analysis of SEM X-ray spectral images: A powerful new microanalysis tool. *Microsc Microanal* **9**, 1-17.

Liebske, C. (2015). iSpectra: An Open Source Toolbox for the analysis of spectral images recorded on scanning electron microscopes. *Microsc Microanal* **21**, 1006-1016.

MALOY, A.K. & TREIMAN, A.H. (2007). Evaluation of image classification routines for determining modal mineralogy of rocks from X-ray maps. Am Mineral 92, 1781–1788.

Mori, N., Kato, N. & Morita, M. (2017). Automatic processing of element maps by automatic colour map filter and high speed cluster analyses for EPMA. EMAS 2017 Conference abstract.

Muir, D. D., Blundy, J.D. & Rust, A.C. (2012). Multiphase petrography of volcanic rocks using element maps: a method applied to Mount St Helens, 1980-2005. *Bull Volcanol* **74**, 1101-1120

Munch, B., Martin, L.H.J. & Leemann, A. (2015). Segmentation of elemental EDS maps by means of multiple clustering combined with phase identification. *J Microsc* **260**, 411-426.

Statham, P., Penman, C., Chaldecott, J., Burgess, S., Sitzman, S. & Hyde, A. (2013). Validating a New Approach to the Mapping of Phases by EDS by Comparison with the Results of Simultaneous Data Collection by EBSD. *Microsc Microanal* **19 (S2)**, 752-753

Tan, P.N., Steinbach, M., Kumar, V. (2006). Introduction to Data Mining. Boston, USA: Pearson Education, Inc.

van den Berg, R.A, Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K. & van der Werf M.J. (2006). Centering, scaling and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142

VAN HOEK, C.J.G., DE ROO, M., VAN DER VEER, G. & VAN DER LAAN, S.R. (2011). A SEM-EDS study of cultural heritage objects with interpretation of constituents and their distribution using PARC data analysis. *Microsc Microanal* **17**, 656–660.

Figure 1. (a) BSE image showing mapped area from which the phase map was generated. Figures 2 and 3 correspond to expanded regions highlighting features within the phase maps. (b) RGB composite image of the first three principle components. Locations from which the area compositions where extracted for each identified phase are shown. (c) BSE Image showing the finely intergrown phase hydrogarnet and serpentine (purple colour on 1b). Location of image is given by white rectangle marked 'c' on 1b. Mineral abbreviations are as follows: Bul - bultfonteinite, Hgt - hydrogarnet and Srp - serpentine.

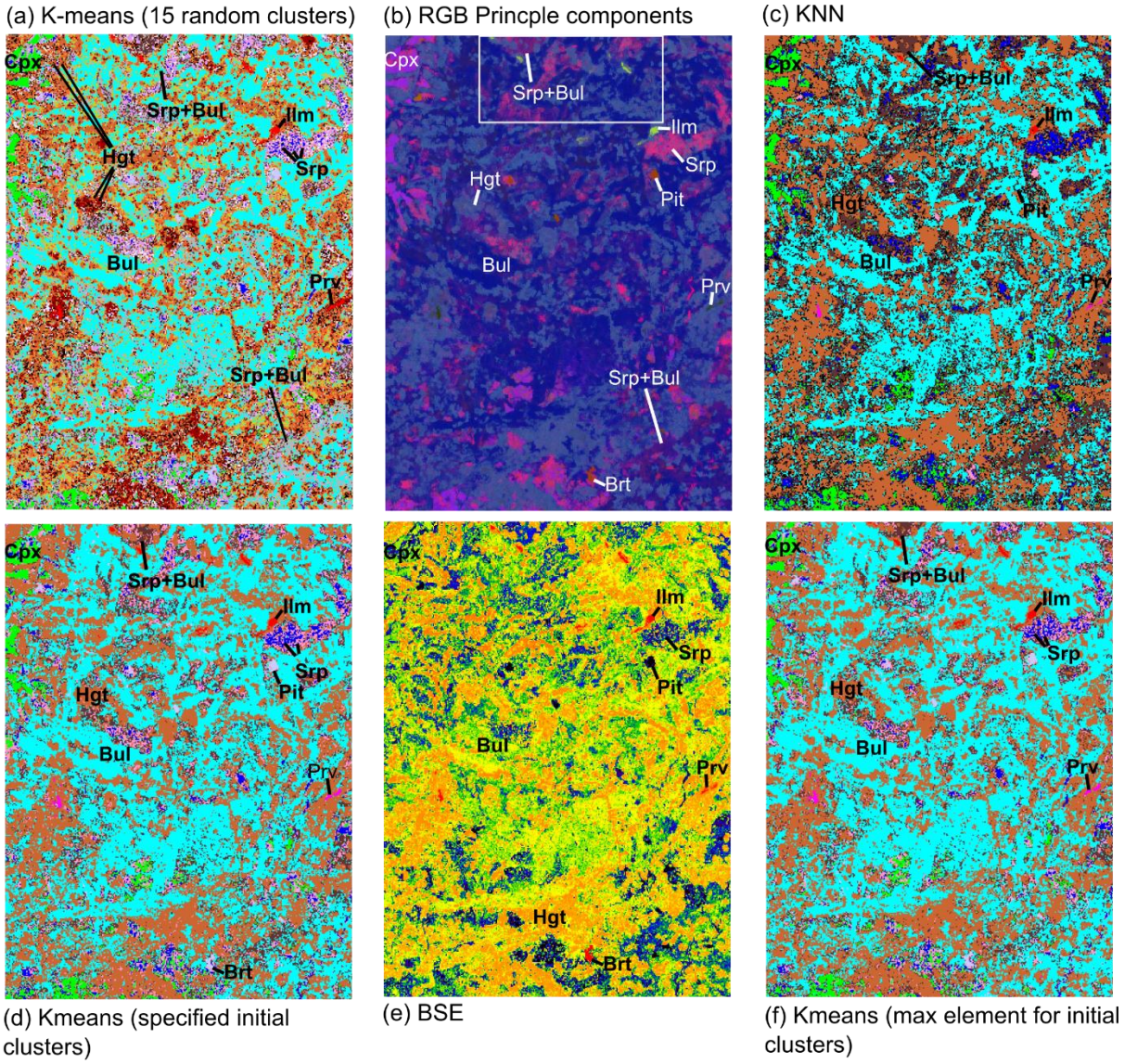
(a) K-means (15 random clusters)
(b) RGB Princple components
(c) KNN
Cpx
Srp+Bul
Ilm
Hgt
Srp
Pit
Bul
Prv
Brt
(d) Kmeans (specified initial clusters)
(e) BSE
(f) Kmeans (max element for initial clusters)
(g) K-means (15 random clusters)
(h) RGB Princple components
(i) KNN
(j) Kmeans (specified initial clusters)
(k) BSE
(l) Kmeans (max element for initial clusters)

Figure 2. Comparison of phase maps with a false-coloured BSE image and RGB-PC image. The area shown is a small region of the map, part of the basalt xenolith. The location is given on Figure 1a. For the KNN algorithm (c) black pixels are unclassified pixels. High magnification images (g-l) are shown for the area represented by the white box in (b). Arrows in (l) correspond to areas where serpentine-bultfonteinite is underrepresented in comparision to (h) and (i). Circles in (l) correspond to misclassified convoluted pixels. Mineral abbreviations are as follows: Brt - barite, Bul - bultfonteinite, Cpx - clinopyroxene, Hgt - hydrogarnet, Ilm - ilmenite, Prv - perovskite and Srp – serpentine.

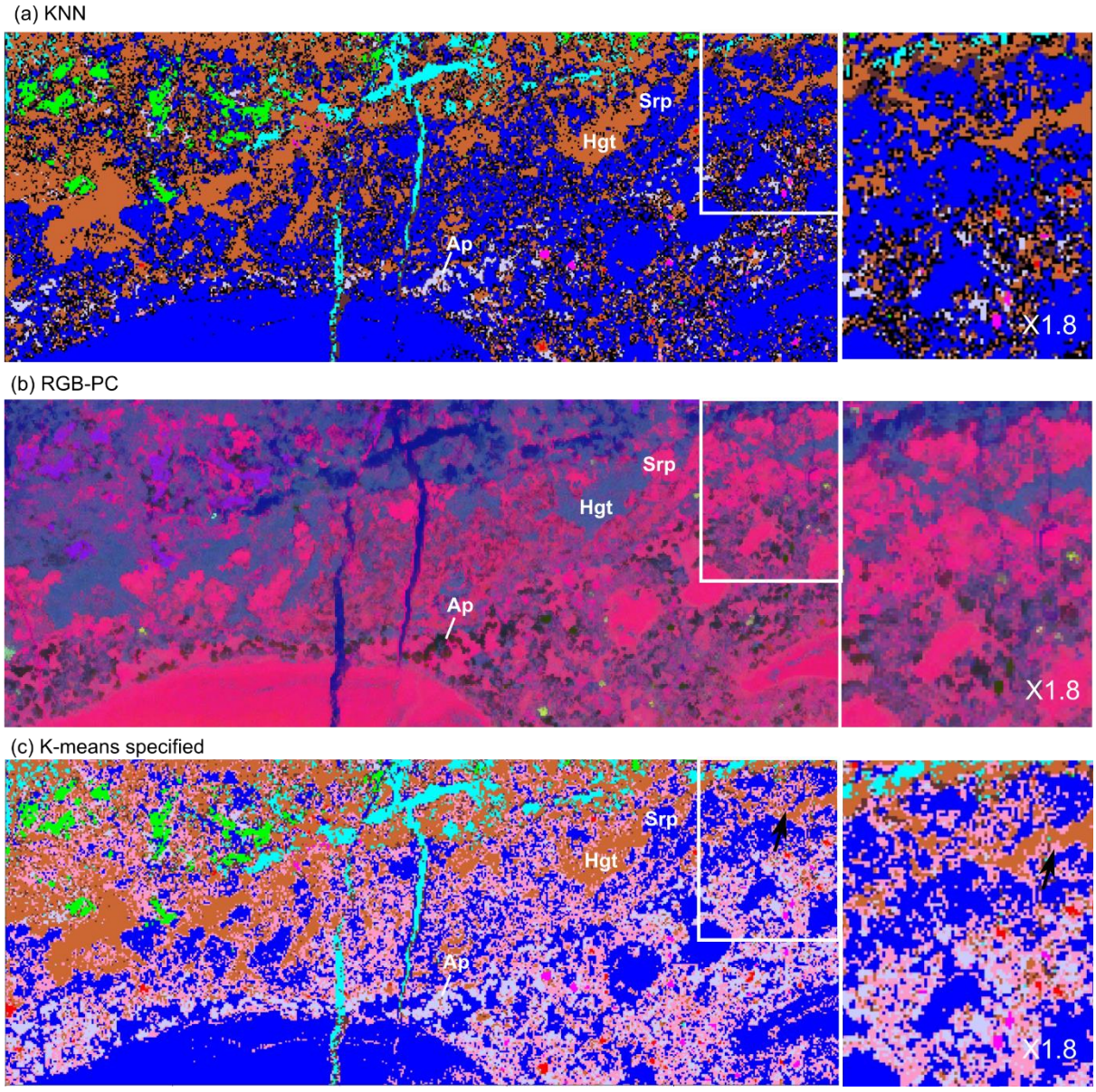

Figure 3. Comparison of phase maps with a RGB-PC image for the kimberlite region of the map. The location is show on Figure 1a. High magnification images are shown to the right for the area represented by the white box. For KNN algorithm (a) black pixels are unclassified pixels.

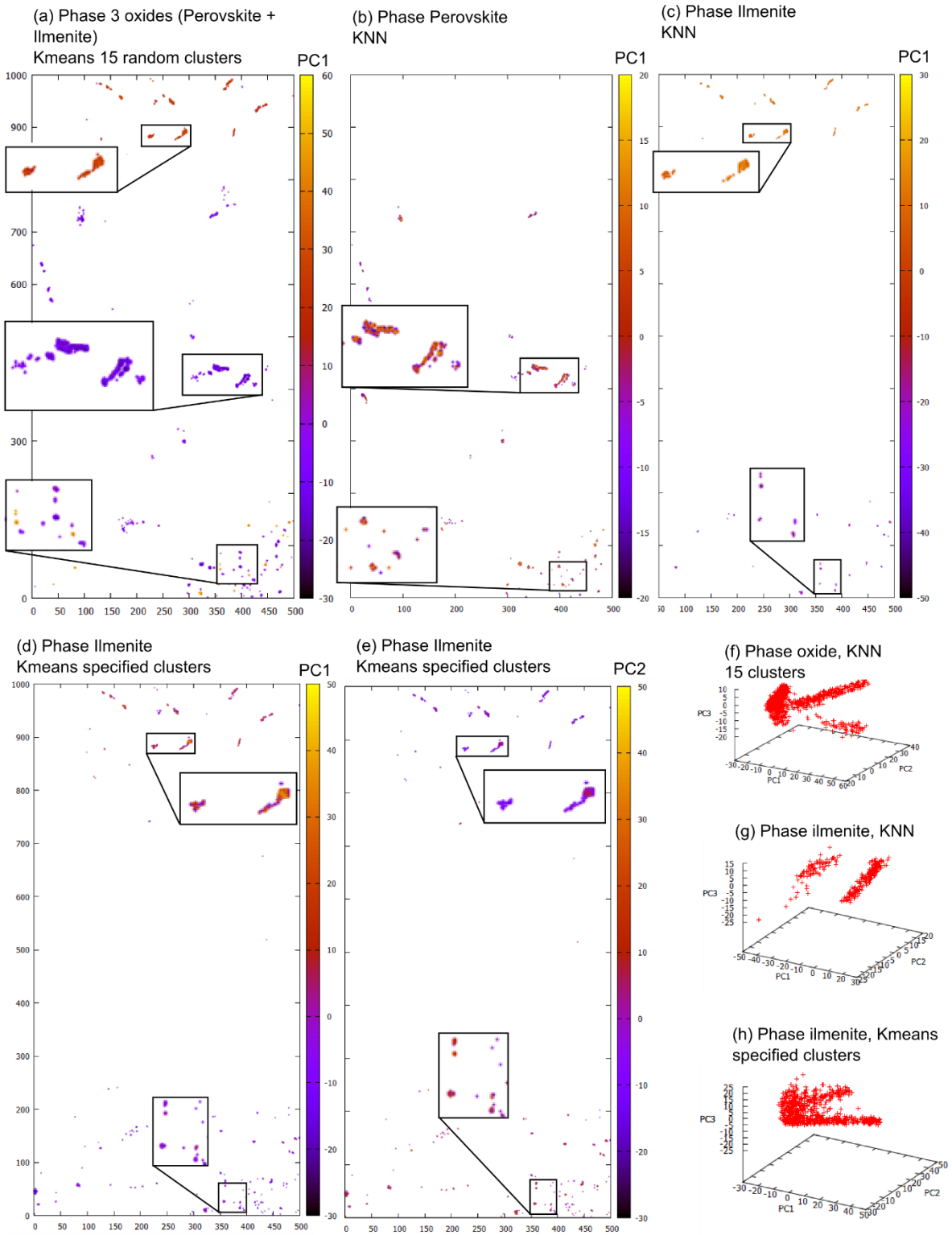

Figure 4. Phase-PC Maps (a-e) examining variation within the oxide, perovskite and ilmenite phases. High magnification insets are shown for the bottom, middle and top of the maps, which correspond to the kimberlite, xenolith margin and xenolith interior. PC scatter graphs (f-h) show the alignment of orthogonal principle components with data variance.

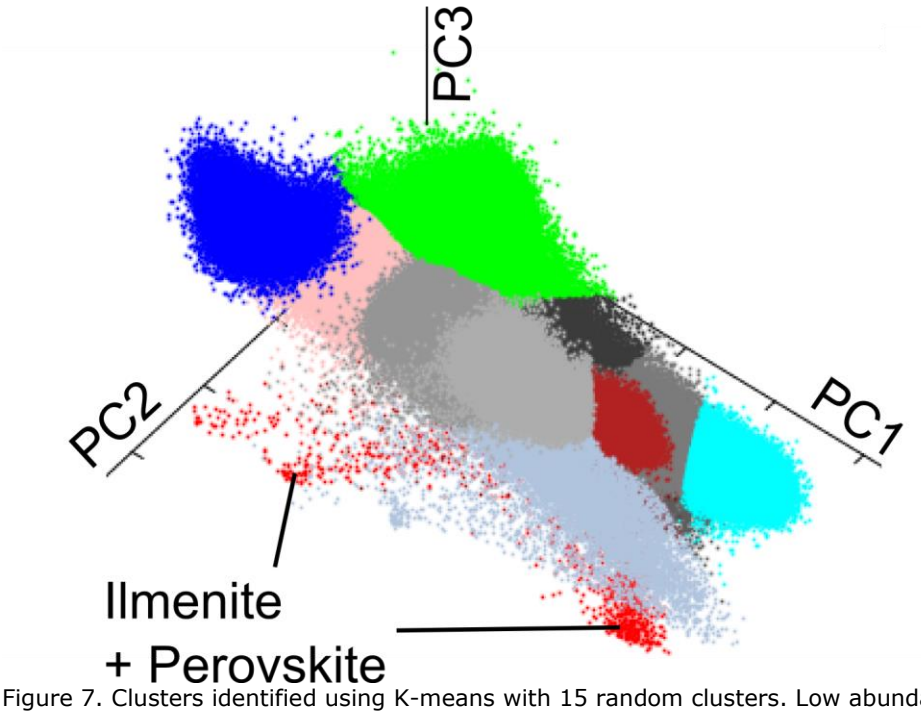

Figure 7. Clusters identified using K-means with 15 random clusters. Low abundance phases ilmenite and perovskite are identified as a single cluster.

Table 1. List of phases identified from RGB composite image of the first three principle components. Numbers correspond to those given on Figure 1b.

Phase	Mineral	Ideal Formula
1	Ilmenite	$FeTiO_3$
2	Pit	
3	Bultfonteinite	$Ca_2SiO_2(OH,F)_4$
4	Clinopyroxene	$Ca(Mg,Fe)Si_2O_6$
5	Serpentine/Chlorite	$(Mg,Fe,Mn,Al)_{12}(SiAl)_8O_{20}(OH)_{16}$
6	Perovskite	$CaTiO_3$
7	Hydrogarnet	$Ca_3(Fe,Ti,Al)_2Si_2O_8(OH)_4$
8	Serpentine + Bultfonteinite intergrowth	
9	Sr-Apatite	$(CaSrBa)_5(PO4)_3(OH,F)$

Table 2. Average phase compositions and standard deviations for K-means clustering using 15 random clusters. Phase identification was made with reference to spatial distribution, BSE and RGB-PC images. High abundance elements are shown in bold. Standard deviation % is given in italic.

Identification	Na2O	MgO	Al2O3	SiO2	K2O	CaO	TiO2	FeO	MnO	Total
Hydrogarnet	0.08	4.06	7.03	**29.97**	0.04	**31.50**	1.36	**13.18**	0.22	87.44
	690.58	*47.43*	*26.51*	***6.37***	*412.50*	***6.20***	*72.00*	***13.76***	*167.54*	
Srp+Bul	0.20	**11.61**	2.12	**32.84**	0.04	**29.15**	0.37	4.32	0.09	80.74
	281.50	***27.15***	*67.94*	***12.60***	*484.84*	***10.38***	*133.59*	*55.83*	*336.37*	
Oxides	0.71	2.63	2.90	8.96	0.05	**26.39**	**36.74**	**12.66**	0.23	91.27
	157.87	*151.71*	*75.29*	*66.45*	*377.52*	***39.64***	***32.41***	***113.57***	*219.36*	
Hydrogarnet	0.11	1.54	2.57	**26.71**	0.01	**40.14**	0.43	5.97	0.11	77.59
	508.51	*84.70*	*55.00*	***9.55***	*899.97*	***4.75***	*98.20*	*33.63*	*292.45*	
Hydrogarnet	0.07	2.01	7.36	**24.95**	0.03	**33.22**	3.07	**11.75**	0.18	82.64
	743.48	*85.81*	*34.04*	***8.50***	*464.74*	***6.44***	*104.75*	***16.55***	*199.97*	
Serpentine	0.16	**23.61**	5.79	**33.52**	0.21	**10.44**	0.46	7.16	0.15	81.51
	310.02	***13.17***	*62.19*	***12.43***	*261.26*	***36.97***	*110.00*	*38.96*	*214.63*	
Hydrogarnet	0.09	2.26	4.56	**27.84**	0.02	**36.47**	0.81	**10.55**	0.19	82.80
	591.64	*83.25*	*35.22*	***7.70***	*552.26*	***5.36***	*74.91*	***18.29***	*190.54*	
Serpentine	0.08	**31.68**	4.69	**36.12**	0.12	3.04	0.13	4.68	0.11	80.65
	521.29	***10.85***	*55.41*	***10.33***	*393.15*	*93.63*	*175.92*	*45.56*	*268.42*	
Augite	0.67	**15.47**	1.45	**49.23**	0.02	**20.79**	0.56	6.48	0.13	94.80
	127.56	***22.36***	*110.15*	***9.87***	*901.96*	***16.23***	*64.61*	*29.59*	*246.93*	
Bul+Srp	0.14	6.39	1.37	**28.78**	0.02	**36.60**	0.19	2.41	0.06	75.95
	379.70	*36.03*	*78.49*	***10.10***	*686.74*	***7.04***	*149.18*	*75.24*	*519.18*	
Hydrogarnet	0.05	1.88	6.86	**27.04**	0.04	**33.36**	1.32	**15.96**	0.23	86.75
	964.25	*82.19*	*25.98*	***7.49***	*416.93*	***6.79***	*82.74*	***13.95***	*161.73*	
Apatite	0.76	7.74	3.32	**14.92**	0.08	**28.55**	0.73	4.19	0.09	60.38
	137.05	*56.73*	*124.80*	***35.46***	*207.64*	***23.65***	*177.01*	*64.07*	*347.44*	
Serpentine	0.18	**16.98**	5.14	**30.61**	0.11	**19.10**	0.83	8.16	0.16	81.27
	317.18	***19.57***	*62.35*	***14.06***	*285.14*	***17.29***	*122.78*	*41.28*	*210.56*	
?	0.10	9.70	6.02	28.58	0.06	26.35	1.42	11.94	0.20	84.38
	529.96	*26.32*	*41.95*	*11.15*	*323.73*	*11.17*	*105.09*	*24.38*	*175.08*	
Bulfonteinite	0.11	1.29	0.64	**27.57**	0.01	**44.12**	0.10	1.16	0.03	75.02
	466.91	*105.34*	*120.89*	***8.64***	*1202.30*	***5.07***	*213.24*	*101.80*	*1110.73*	

Table 3. Comparison of phase average and extracted composition for a selected area. Data is for the clinopyroxene phase, with phase average calculated from K-means clustering using 15 random clusters. High abundance elements are shown in bold. Standard deviation % is given in italic. Atoms per formula unit (apfu) are calculated on the basis of 6 oxygens.

Wt. %	**Na2O**	**MgO**	**Al2O3**	**SiO2**	**K2O**	**CaO**	**TiO2**	**FeO**	**MnO**	**Total**
Phase average	0.67 *127.56*	**15.47** ***22.36***	1.45 *110.15*	**49.23** ***9.87***	0.02 *901.96*	**20.79** ***16.23***	0.56 *64.61*	6.48 *29.59*	0.13 *246.93*	94.80
Spot extract	0.18 *410.93*	**15.40** ***9.22***	0.18 *226.55*	**55.16** ***3.52***	-0.04 *279.49*	**23.49** ***9.70***	**0.30** ***55.15***	4.70 *33.00*	0.19 *170.59*	**99.57** ***3.85***
apfu	**Na**	**Mg**	**Al**	**Si**	**K**	**Ca**	**Ti**	**Fe**	**Mn**	**Total**
Phase average	0.051	0.902	0.067	1.926	0.001	0.871	0.016	0.212	0.004	4.050
Spot extract	0.013	0.843	0.008	2.024	0.000	0.924	0.008	0.144	0.006	3.970

Table 4 Extracted compositions for the discrete groupings identified from phase-PC1 map of ilmenite phase (KNN algorithm). For comparison the extracted composition for ilmenite used in the reference dataset is given. Standard deviation are given in italic.

PC1 threshold (location)	**Na2O**	**MgO**	**Al2O3**	**SiO2**	**K2O**	**CaO**	**TiO2**	**FeO**	**MnO**	**Identification**
<0 (In kimberlite)	0.00 *0.67*	14.01 *4.60*	4.05 *1.80*	3.89 *3.32*	0.00 *0.17*	4.68 *3.36*	16.42 *4.61*	48.37 *8.09*	0.92 *0.57*	Spinel
>0 (In xenolith)	0.00 *0.91*	2.13 *1.23*	2.25 *1.84*	7.70 *4.66*	0.00 *0.15*	9.45 *5.55*	39.58 *7.59*	33.62 *5.71*	0.71 *0.55*	ILM
reference composition (from xenolith)	1.97	3.43	0.00	0.00	0.00	0.92	46.93	43.15	0.10	ILM

Table 5 Extracted compositions for the discrete groupings of serpentine identified in the phase-PC map. Olivine pseudomorphs are the large sub-rounded grains within the kimberlite which were original olivine but have been replaced by serpentine.

PC1 Threshold (location)	Na2O	MgO	Al2O3	SiO2	K2O	CaO	TiO2	FeO	MnO	Total
> 0 (xenolith)	0.08	33.20	3.35	38.86	0.10	3.89	0.15	4.02	0.08	83.74
< -4 (Kimberlite: rims of olivine pseudomorphs	0.09	28.79	7.56	32.05	0.21	2.12	0.12	6.25	0.17	77.36
<0 and >-4 (Kimberlite: olivine pseudomorphs)	0.07	30.29	5.91	33.60	0.15	2.26	0.11	5.29	0.14	77.81